Linguistique informatique
Traduction automatique

Laurence Danlos

Introduction

Dans le célèbre film "2001, L'Odyssée de l'espace" de S. Kubrik, l'ordinateur HAL comprend l'homme, dialogue avec lui dans sa langue, exécute ses commandes, et ressent des émotions. En 1968, Marvin Minsky, conseiller scientifique du film en tant que spécialiste de L'Intelligence Artificielle, pensait qu'on pourrait effectivement réaliser un ordinateur tel que HAL en 2001. En l'an 2000, est-on prêt de la réalisation de HAL ? Cet article essayera de montrer et d'expliquer pourquoi on en est loin.

Cette affirmation peut paraître d'emblée critiquable à l'heure où les applications de la linguistique informatique (ou plus généralement de l'Intelligence Artificielle) envahissent à un rythme accéléré notre quotidien. En effet, bon nombre de voitures parlent, tout traitement de textes intègre un correcteur orthographique, des systèmes de dictée automatique ou de traduction automatique fleurissent sur le marché, on nous annonce régulièrement des photocopieurs ou téléphones qui traduisent, les moteurs de recherche sur le Web proposent un résumé et une traduction des textes trouvés, on peut bénéficier d'une assistante virtuelle personnelle pour gérer nos rendez-vous, etc. Bref, des applications que l'on n'aurait même pas imaginées il y a quatre ou cinq ans existent bel et bien aujourd'hui. Pourquoi est-on donc si loin de HAL ? La réponse à cette question repose sur une notion clef : la **compréhension**. La réalisation de HAL demande que l'ordinateur comprenne le langage, tandis que les applications que l'on propose aujourd'hui ne supposent pas de compréhension ou alors une compréhension très superficielle. Et c'est là où se situe la différence : on ne sait pas simuler à l'heure actuelle la compréhension du langage sur ordinateur.

Cet article sera divisé en deux parties. Dans une première partie, nous présenterons des applications où l'ordinateur manipule du langage écrit ou oral sans rien y comprendre ou pas grand chose. Nous présenterons par là même ce que peut être la compréhension pour un ordinateur et discuterons

de "l'âge mental des robots humanoïdes". Dans la seconde partie, nous présenterons une application particulière, la traduction automatique, où l'ordinateur aurait besoin de comprendre : on ne traduit pas sans comprendre. A travers cette application, nous exposerons quelques grandes difficultés rencontrées dans le traitement automatique du langage écrit. Ces difficultés expliquent les erreurs fréquemment observées dans les textes traduits automatiquement et illustrent la distance à parcourir pour simuler la compréhension humaine sur ordinateur.

1 Applications sans compréhension ou avec une compréhension superficielle

Nous allons présenter deux applications, l'une réalisée en 1964, l'autre de nos jours. Ceci nous permettra de mesurer la distance parcourue et la distance qui reste à parcourir.

Dès 1964, J. Weizembaum et ses collègues de MIT ont réalisé un logiciel, Eliza, qui a impressionné et impressionne encore. Ce logiciel propose une simulation de dialogues écrits entre un psychothérapeute et un patient, le rôle du psychothérapeute étant tenu par l'ordinateur. Donnons tout de suite un exemple de dialogue : les répliques du patient saisies sur l'ordinateur par un être humain sont en italiques, celles du psychothérapeute en gras.

Bonjour, je suis votre psychothérapeute. Parlez-moi de vos problèmes.
Je suis très angoissé.
Pourquoi dites-vous que vous êtes très angoissé ?
C'est à cause de ma femme.
Parlez-moi de votre famille.
J'adore mes enfants.
J'aimerais que vous me parliez de vos relations avec vos enfants.
Personne ne m'aime.
Est-ce que ça vous ennuie que personne ne vous aime ?
Oui, je me sens seul.
Votre vie sexuelle y est peut-être pour quelque chose. Continuez sans crainte.
...

On s'y croirait ! Le dialogue est époustouflant ! Et pourtant l'ordinateur psychothérapeute ne comprend rien, absolument rien aux répliques de son patient. Comment cela est-il possible ? Par la technique du "pattern-

matching" que nous allons expliquer brièvement sur les échanges de notre dialogue.

Bonjour, je suis votre psychothérapeute. Parlez-moi de vos problèmes.

C'est un des textes pré-enregistrés par lequel l'ordinateur psychothérapeute entame son dialogue. Ce texte est choisi aléatoirement parmi un ensemble de textes pré-enregistrés au contenu similaire.

Je suis très angoissé.

C'est une réponse libre du patient qui est stockée dans une variable R1.

Pourquoi dites-vous que vous êtes très angoissé ?

Cette question du psychothérapeute n'est qu'un ajustement morpho-syntaxique de la chaîne de caractères : Pourquoi dites-vous que R1 ? L'ordinateur n'a pas besoin de comprendre le sens de R1 pour poser cette question. Il se contente de transformer **je suis** en **vous êtes**. Si le patient avait dit *Je suis très euphorique*, le "psychothérapeute" aurait tout aussi imperturbablement demandé **Pourquoi dites-vous que vous êtes très euphorique ?**

C'est à cause de ma femme.

C'est une réponse libre du patient qui contient le mot "femme".

Parlez-moi de votre famille.

C'est un ordre systématique du psychothérapeute dès que la réplique du patient contient un mot de la liste : *femme, mari, père, mère,* etc.

Oui, je me sens seul.

Réponse libre du patient qui ne contient rien de spécial.

Votre vie sexuelle y est peut-être pour quelque chose. Continuez sans crainte.

Réponse du psychothérapeute quand il ne sait plus trop quoi dire.

En résumé, l'ordinateur psychothérapeute ne comprend rien aux interventions de son patient : il se contente de produire des réponses en activant une des centaines ou milliers de réponses pré-enregistrées et en effectuant des transformations morpho-syntaxiques (e.g. *je suis* → *vous êtes*).

Près de 40 ans après Eliza, J. Cassel et ses collègues de MIT sont en train de réaliser un logiciel, Rea, qui est à la pointe de la recherche et de la technologie. Ce logiciel propose une simulation de dialogues oraux entre un agent immobilier et un client. Le rôle de l'agent immobilier est tenu par un robot humanoïde, c'est-à-dire un robot à forme humaine capable de communiquer par gestes, par le regard et par la parole (ce qui est connu sous le terme de "communication multi-modale"). Les progrès entre Eliza (1964) et Rea (2000) sont considérables : on est passé d'un dialogue écrit à un

dialogue oral (ce qui est dû aux progrès énormes de la reconnaissance et de la synthèse de la parole), on est passé d'un ordinateur à un humanoïde (ce qui est dû aux progrès énormes de la robotique), et enfin on est passé à une communication multi-modale grâce à une bonne intégration de différentes technologies. Qu'en est-il de la compréhension ? Les progrès sont minces : Rea comprend à peu près ce dont parle son client, mais uniquement si celui-ci se cantonne à des questions concernant l'immobilier. Si le client passe de l'achat d'un appartement à l'achat d'une voiture, Rea est complètement perdue. Ceci est du au fait que le module de compréhension n'a que des connaissances linguistiques ou extra-linguistiques sur l'immobilier. En particulier, son vocabulaire se limite à celui de l'immobilier. D'une manière plus générale, il n'existe aucun système de compréhension générique, c'est-à-dire fonctionnant pour la conversation courante et pour les multiples domaines fermés (immobilier, juridique, médecine, etc.). A l'heure actuelle, l'ordinateur ne peut comprendre un texte (i.e. calculer une représentation sémantique de ce texte suffisamment abstraite pour pouvoir effectuer du raisonnement dessus) que si celui-ci relève d'un domaine fermé et donc restreint linguistiquement et conceptuellement. Les raisons techniques de ces limitations seront expliquées dans la seconde partie de l'article.

Ces limitations dans la compréhension des humanoïdes sont souvent traduites dans les médias dans les termes suivants : "Rea a un âge mental de trois ans". Nous pensons que ce type d'affirmation est fondamentalement erroné. Non pas du fait que l'âge mental est de deux ou quatre ans au lieu de trois, mais simplement du fait qu'il est spécieux de faire une comparaison entre l'âge mental d'un robot humanoïde et celui d'un enfant. En effet, il n'existe à notre connaissance aucun enfant, même atteint de troubles langagiers, qui ne soit capable de comprendre le langage que dans un domaine restreint (e.g. la nourriture) à l'exclusion de tout autre domaine (e.g. les jouets, les câlins, etc.). Autrement dit, on ne peut pas parler de l'âge mental d'un humanoïde en effectuant une comparaison avec l'âge mental d'un enfant. L'apprentissage du langage (et de la perception du monde) par un enfant passe par des mécanismes encore mal identifiés à l'heure actuelle mais qui n'ont forcément rien à voir avec les mécanismes utilisés dans la réalisation d'un humanoïde comme Rea dont on restreint sciemment la connaissance à l'immobilier sans chercher le moins du monde à élargir le champ de ses compétences.

Il est aussi fréquent d'entendre parler des "états d'âme" des robots humanoïdes. Ainsi, un humanoïde "vous confie ses états d'âme", raconte-t-on, lorsqu'il dit : *Mes batteries sont à plat*. Mais alors votre voiture vous confie aussi ses états d'âme lorsque s'allume la lumière rouge de la batterie. Le principe est le même, seuls le mode de communication et la forme de

l'objet diffèrent. Suffit-il d'être un objet à forme humaine et à synthèse vocale pour avoir des états d'âme ?

2 La traduction automatique

Dans un système de traduction automatique (désormais TA), un texte en langue source (notée Ls, le français par exemple) est donné en entrée sous forme électronique au système de TA qui calcule le texte en langue cible (notée Lc, l'anglais par exemple). Cette traduction, qui se présente sous forme électronique, est prête à être imprimée ou diffusée sur le Web.

La TA est l'application de linguistique informatique la plus ancienne : les recherches en TA sont contemporaines des débuts mêmes de l'informatique (vers la fin des années 1940). C'est une application très prisée car les besoins sont énormes. Ainsi la Commission Européenne traduit environ un million de pages par an, les multinationales environ un milliard.

Les premiers systèmes de TA reposaient sur une traduction mot à mot qui est schématisée dans la Figure 1. La phase de "lemmatisation" en Ls consiste à mettre les noms au singulier, les verbes à l'infinitif, etc. La phase de transfert consiste à associer à un mot en Ls sa traduction en Lc (e.g. *noir* → *black*). La phase de réajustement en Lc permet de respecter les règles morpho-syntaxiques de Lc (e.g un adjectif se place avant le nom en anglais).

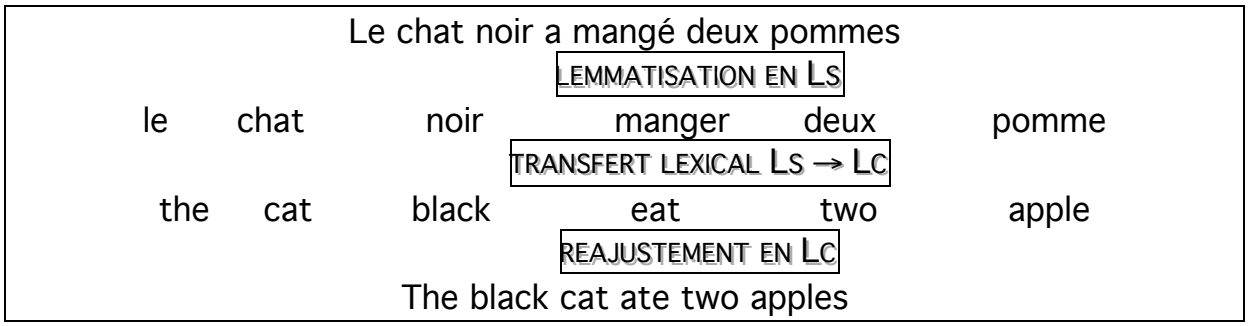

Figure 1: Traduction mot à mot

On ne connaît que trop bien les limites de la traduction mot à mot qui donne pour la phrase (1) ci-dessous la traduction erronée (2) au lieu de la traduction correcte (3).

(1) Un pied noir a mangé une pomme de terre
(2) A black foot ate an apple of earth
(3) An Algerian-born Frenchman ate a patato

La traduction mot à mot est vouée à l'échec car il est connu que l'on ne peut pas traduire sans comprendre. Un système de TA doit donc (en principe) comporter un module d'analyse et un module de génération, voir Figure 2. Le module d'analyse est chargé de la compréhension du texte en Ls et calcule une représentation sémantique de ce texte. Celle-ci est fournie en entrée au module de génération qui produit le texte en Lc.

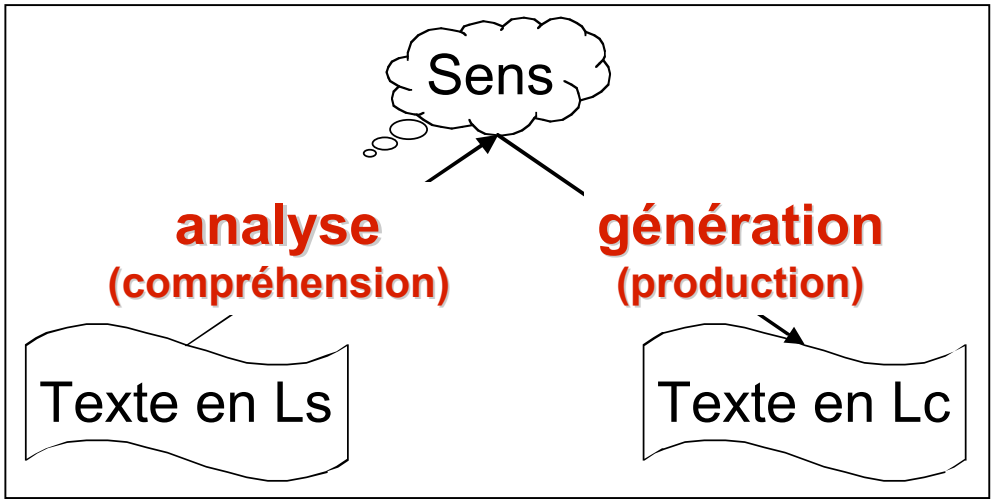

Figure 2 : Architecture d'un système de TA

Mais cette architecture d'un système de TA relève plus du principe que de la réalité car on n'arrive guère à réaliser des modules d'analyse ou de génération. Pourquoi ? Parce que la langue comporte une infinité d'ambiguïtés. La relation sens-forme n'a rien de biunivoque : d'une part, une forme linguistique donnée a plusieurs sens (ce qui crée des ambiguïtés en analyse), d'autre part un sens donné peut s'exprimer par plusieurs sens (ce qui crée des ambiguïtés en génération). Nous allons illustrer les ambiguïtés de la langue uniquement en analyse et uniquement sur le cas des homographes, c'est-à-dire sur deux mots qui ont la même graphie mais des sens différents (et donc généralement des traductions différentes). Ainsi, *le* est soit un article soit un pronom, *savoir* est soit un verbe soit un nom, *avocat* est toujours un nom mais soit il désigne un *homme de loi* (et se traduit alors par *lawyer*) soit il désigne un *fruit* (et se traduit alors par *avocado*). Dans la plupart des cas, les deux sens de *avocat* ne crée pas d'**ambiguïté réelle** : dans les phrases (4) et (5) ci-dessous un humain désambiguïse facilement ce mot par le contexte où il est employé et établit que *avocat* désigne un *fruit* en (4) et un *homme de loi* en (5).

(4) Zoé a mangé un avocat
(5) Zoé a rendez-vous avec un avocat

Mais pour l'ordinateur il y a une **ambiguïté virtuelle** introduite par le module d'analyse qui connaît les deux sens de *avocat*. Cette ambiguïté

virtuelle doit impérativement être levée pour éviter des traductions erronées comme (6) et (7).

(6) Zoé ate a lawyer
(7) Zoé has a meeting with an avocado

La levée des ambiguïtés virtuelles pour les homographes demande d'effectuer deux tâches :

établir une classification sémantique des noms (les classes de noms sont écrites en majuscules), par exemple :

avocat est un FRUIT qui est un COMESTIBLE
avocat est un HUMAIN

établir la catégorie sémantique des compléments de verbe, par exemple :

ANIME manger COMESTIBLE
HUMAIN avoir rendez-vous avec HUMAIN

Mais l'affaire se complique car les verbes sont aussi souvent des homographes. Ainsi, *manger* a les sens illustrés en (4) et dans les exemples suivants :

(8) Ce poêle mange beaucoup de charbon
APPAREIL manger COMBUSTIBLE
This stove uses a lot of coal

(9) Les grosses entreprises mangent les petites
ORGANISATION manger ORGANISATION
Big firms swallow up smaller ones

(4) présente donc un phénomène d'ambiguïté croisée : pour désambiguïser *avocat*, il faut désambiguïser *manger* et pour désambiguïser *manger*, il faut désambiguïser *avocat*. Les ambiguïtés croisées donnent lieu à une explosion combinatoire que l'on peut chiffrer ainsi : si une phrase a n mots m_1 m_2 ... m_i ... m_n et si un mot m_i a k_i sens (donc généralement k_i traductions), alors l'ordinateur doit choisir entre K hypothèses avec $K = k_1 \times k_2 \times \ldots \times k_i \times \ldots \times k_n$. On dit que le module d'analyse surgénère en produisant une prolifération d'hypothèses. Il arrive de plus qu'un homographe ne puisse pas être désambiguïsé par le contexte immédiat (la phrase où il apparaît) ni par un contexte plus large. Ainsi, en (10), on ne peut désambiguïser *avocat* ni par *aimer*, ni par *véreux*, ces deux prédicats s'appliquant tant à des fruits qu'à des humains.

(10) Zoé a aimé cet avocat. Pourtant, il était véreux.
Zoé loved / liked this lawyer / avocado. However, he / it was shady / worm-eaten.

Soulignons le point suivant : le texte en (10) est réellement ambigu, mais il ne sera que rarement perçu comme tel en situation d'énonciation, par exemple dans un dialogue entre deux personnes car celles-ci savent bien si elles sont en train de parler des amours de Zoé ou de ce qu'elle a mangé à midi. En TA, ou plus généralement en compréhension, on peut simuler ce type de connaissances en restreignant le domaine des textes à traiter : un système comprenant un module d'analyse ne peut produire des résultats satisfaisants que dans un domaine fermé comme le juridique, l'informatique ou l'immobilier. En effet, un module d'analyse dédié au juridique peut ne retenir que le sens *homme de loi* du mot *avocat* et ainsi limiter la prolifération d'hypothèses dues aux homographes. Cette méthode, qui n'est pas sans induire quelques erreurs, est un point de passage obligé car, rappelons-le, les homographes ne sont qu'une illustration des ambiguïtés de la langue : c'est une ambiguïté de type sémantique mais il existe d'autres ambiguïtés sémantiques (e.g. les temps des verbes, par exemple un présent utilisé pour un événement futur comme dans *Zoé vient demain*) et des ambiguïtés à tous les niveaux de la langue : morphologique, syntaxique et pragmatique. Cette foison d'ambiguïtés, qui débouche pour un texte sur un réseau complexe d'ambiguïtés croisées, ne peut être gérée en grandeur réelle, c'est-à-dire en simulant la compréhension humaine capable d'appréhender tant les conversations courantes que les conversations plus techniques.

Il n'en reste pas moins que les systèmes de traduction automatique progressent lentement mais sûrement avec deux tendances. Dans les laboratoires de recherche, des prototypes sont développés qui donnent des résultats assez satisfaisants dans des domaines restreints et moyennant l'utilisation d'ordinateurs puissants et un temps de calcul conséquent. Sur le marché, les produits commerciaux fleurissent. Ils produisent souvent des traductions erronées mais qui peuvent au moins servir à déterminer de quoi parle le texte (ce qui est important en vieille technologique, par exemple). Ces produits couvrent un large vocabulaire, fonctionnent sur des PC et produisent instantanément la traduction. On peut penser que ces deux tendances vont converger avec le développement de la puissance des ordinateurs. Certes, mais ceci n'est pas suffisant. Il reste un énorme travail à effectuer qui consiste à formaliser et à enregistrer dans l'ordinateur toutes les connaissances linguistiques et extra-linguistiques, et ceci ne sera pas achevé en 2001 !

Remerciements
Je remercie mes nombreux amis et collègues qui m'ont aidé pour cette présentation, plus particulièrement Isabelle Faugeras, Frédéric Meunier, Fayez Okdeh, Gaëlle Récourcé et Laurent Roussarie.